Matching Pursuit Shrinkage in Hilbert Spaces

Tieyong Zeng, François Malgouyres

Matching Pursuit Shrinkage in Hilbert Spaces

Tieyong Zeng[1*] and François Malgouyres[2]

Abstract

In this paper, we study a variant of the Matching Pursuit named Matching Pursuit Shrinkage. Similarly to the Matching Pursuit it seeks for an approximation of a datum living in a Hilbert space by a sparse linear expansion in an enumerable set of atoms. The difference with the usual Matching Pursuit is that, once an atom has been selected, we do not erase all the information along the direction of this atom. Doing so, we can evolve slowly along that direction. The goal is to attenuate the negative impact of bad atom selections.

We analyse the link between the shrinkage function used by the algorithm and the fact that the result belongs to an l^p space.

Index Terms

Dictionary, Matching Pursuit, shrinkage, sparse representation.

EDICS Category: SMR-REP, TEC-RST

[1]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong. e-mail: zeng@hkbu.edu.hk, tel: (+852) 3411 2531, fax: (+852) 3411 5811.

[2]Université Paris 13, CNRS UMR 7539 LAGA, 99 avenue Jean-Batiste Clément, F-93 430 Villetaneuse, France, e-mail: malgouy@math.univ-paris13.fr, tel: (+33) 149 403 583, fax: (+33) 149 403 568.

[*]Correspondence author

I. Introduction

A. Recollection on sparse approximation

Finding a sparse approximation of a data in a Hilbert space is a reccurent problem in applied science. The problem is to approximate a datum $v \in \mathcal{H}$ ($\mathcal{H}$ is a Hilbert space of finite or infinite dimension) by a linear expansion in a dictionary of known atoms $(\psi_i)_{i\in I}$:

$$v \sim \sum_{i\in I} \lambda_i \psi_i,$$

where $(\lambda_i)_{i\in I} \in \mathbb{R}^I$. The approximation is needed because v is usually corrupted by noise. Also, it is sometimes preferable to search for an approximation which is coarser than the noise requires. Doing so we favors desired/expected properties of the coordinates $(\lambda_i)_{i\in I}$.

Moreover, the dictionary is usually overcomplete. This offers the freedom to select among all the possible sets of coordinates one of those agreeing with some prior knowledge or desired property of the coordinates. The property receiving most of the attention is sparsity. Heuristically, we select the set of coordinates offering the "simplest" explanation of the datum. Rigorously, for a given accuracy after reconstruction, we want

$$l^0\left((\lambda_i)_{i\in I}\right) \stackrel{\text{def}}{=} \#\{i \in I, \lambda_i \neq 0\},$$

to be as small as possible, where $\#$ denotes the cardinality of a set.

Unfortunately, problems similar to

$$\begin{cases} \text{minimize } l^0\left((\lambda_i)_{i\in I}\right) \\ \text{under the constraint } \| \sum_{i\in I} \lambda_i \psi_i - v\| \leq \tau \end{cases} \tag{1}$$

where $\tau > 0$ and $\|.\|$ is the norm associated with the scalar product of the considered Hilbert space, are known to be NP-Hard in general (see [1]).

As a conclusion, solving (1) is both an open and interesting problem. It receives a lot of attention and it is impossible to list all the contributions to its resolution. Before describing the most popular technics, we give in the next section the algorithm studied in this paper. It will then be simpler to motivate our proposal.

B. The Matching Pursuit Shrinkage

The Matching Pursuit Shrinkage (MPS) is very similar to the usual Matching Pursuit (MP) algorithm (see [2]). The main difference is that it uses a shrinkage[1] function $\theta : \mathbb{R} \to \mathbb{R}$. We describe the algorithm

[1]The rigorous definition of shrinkage functions is given in Section II.

in Table I.

- Input : A datum v, a dictionary $(\psi_i)_{i\in I}$, a shrinkage function θ and $\alpha \in [0,1]$
- Output : Coordinates $(s_n, \gamma_n)_{n\in\mathbb{N}}$
- The algorithm
  - Initialize $R^0 v = v$
  - Repeat until convergence (loop in n)
    1) Select a well correlated atom ψ_{γ_n} such that
    $$|\langle \psi_{\gamma_n}, R^n v\rangle| \geq \alpha \sup_{i\in I} |\langle R^n v, \psi_i\rangle|; \tag{2}$$
    2) Evolve along ψ_{γ_n}
    $$R^n v = s_n \psi_{\gamma_n} + R^{n+1} v, \tag{3}$$
    where
    $$s_n = \theta(M_n) \text{ with } M_n = \langle R^n v, \psi_{\gamma_n}\rangle. \tag{4}$$

TABLE I

THE MATCHING PURSUIT SHRINKAGE (MPS).

Several convergence criterion might be considered but, for simplicity, we always assume that the algorithm stops whenever $s_n = 0$.

Whenever they exist, we can construct coordinates

$$\lambda_i = \sum_{n\in\mathbb{N}:\gamma_n=i} s_n, \qquad \forall i \in I \tag{5}$$

from the result of the MPS. We also consider (when they exist)

$$u = \sum_{i\in I} \lambda_i \psi_i = \sum_{n=0}^{+\infty} s_n \psi_{\gamma_n}.$$

Notice that if we sum (3) for $n = 0 \dots N-1$, we obtain

$$v = \sum_{n=0}^{N-1} s_n \psi_{\gamma_n} + R^N v. \tag{6}$$

This explains the name "residual error" for $R^N v$.

C. Other algorithms promoting sparsity

One of the oldest and simplest algorithm for building a sparse approximation is the Matching Pursuit (MP) [2] or Projection Pursuit [3]. It corresponds to the algorithm of Table I when θ is the identity (i.e. $s_n = M_n$).

In finite dimension (see [2]) and in infinite dimension but under restrictive conditions on the dictionary and the signal (see [4]), the MP is known to converge exponentially. When no hypotheses are made on the dictionary, we only know that the MP converges (see [2]). Some examples show that we cannot expect a "good" converge rate in the most general setting (see [5]). Though the MP and the best k-term approximation have a similar convergence, when the dictionary is "quasi-orthogonal" (see [6]).

There exists "fast" variants of the MP (see [7]). Also, a real-time implementation of the MP is available for audio signal processing (see [8]). The improved performance are obtained by carefully optimizing the structures, algorithms and their implementation. In particular, the update of $(\langle R^n v, \psi_i \rangle)_{i \in I}$ and the computation of γ_n satisfying (2) (in Table I) are implemented in a very efficient way. Each iteration of the MP is typically of complexity $O(log(\#I))$. These optimization are possible because one coordinate only is updated. If K coordinates are modified at each iteration, we obtain a complexity $O(K+log(\#I))$. This might be less favorable when K is large. Althought its approximation performances are not as good as most modern models/algorithms, these acceleration make the MP a usefull algorithm.

The accelerations decribed in [8] can be applied to the MPS, as described in Table I. The potential advantage of introducing a shrinkage function θ is to attenuate the mistakes in the selection of a coordinate γ_n. Let us underline that avoiding wrong selection of coordinates is one of the key ingredient of modern variants of the MP such as CoSaMP [9], Subspace Pursuit [10] and Iterative Hard Thresholding [11]. However, especially when the solution we are looking for is moderately sparse, those algorithms are more computationaly intensive.

Let us go back in time. The most famous variant of the MP is the Orthogonal Matching Pursuit (OMP) (see [12]). In Table I, it replaces the update rule (3) by an orthogonal projection onto the subspace generated by the selected atoms. It is known to provide sparser solutions than the regular MP. From the computational point of view, it has two drawbacks. Firstly, although several attemps have been made to optimize it (see [13], [14]), the orthogonal projection is computationaly expensive and often requires too much memory. Secondly, every selected coordinate is modified. As a consequence, the adaptation of the optimization performed in [8] would only be efficient when the result is very sparse. Algorithms such as the Gradient Pursuit (see [15]) approximately solve the OMP at a cost more similar to the cost of the MP. However, at each iteration, they typically update all the selected coordinates. The computational cost of the Gradient Pursuit is therefore more important than the cost of a fast implementation of the MP, when the solution is moderetely sparse.

Finally, the l^1 regularization (also named Basis Pursuit and Basis Pursuit Denoising, see [16] and the

papers citing it) is a very important sparsity promoting model. It consists in minimizing

$$\|v - \sum_{i \in I} \lambda_i \psi_i\|^2 + \beta \sum_{i \in I} |\lambda_i|$$

and it is very efficient for providing sparse approximations of $v \in \mathcal{H}$. However, its resolution remains (and will probably remain in a near futur) a challenge for large scale problems. A famous (and representative) solver of the l^1 regularization problem is the Iterative Soft Thresholding (see [17]). It updates all the coordinates at each iteration and often requires many iterations before it reaches a suitable convergence level. It is interesting to notice that, in this context, the impact of the choice of the shrinkage function is well understood (see [18]): Every proximal threholding function corresponds to a different objective function.

Inspired by the l^1 regularization problem, a "coordinatewise optimization algorithms" has been proposed in [19]. It performs a soft thresholding, sequencially on each coordinate. The "greedy coordinate descent" proposed in [20] is similar but selects the coordinates according to a criteria similar to the MP. Because they only update one coordinate at each iteration, these algorithms can benefit from the optimization proposed in [8].

D. Notations

The following notations and hypotheses hold all along the paper.

The datum v belongs to a Hilbert space $\mathcal{H}$. The space $\mathcal{H}$ might be of finite or infinite dimension. For any two elements u and w in $\mathcal{H}$, their scalar product is denoted by $\langle u, w \rangle$. As usual, the norm of $u \in \mathcal{H}$ is defined by $\|u\| \stackrel{\text{def}}{=} \sqrt{\langle u, u \rangle}$. The dictionnary $(\psi_i)_{i \in I}$ is made of atoms $\psi_i \in \mathcal{H}$, such that $\|\psi_i\| = 1$, for all $i \in I$. We sometimes denote the dictionary by $\mathcal{D}$. For simplicity, we assume that I is enumerable. In particular, the supremum in (2) may not be reached. In such a case, the MPS is only defined for $\alpha < 1$. For any $u \in \mathcal{H}$, we denote $\|u\|_{\mathcal{D}} \stackrel{\text{def}}{=} \sup_{i \in I} |\langle u, \psi_i \rangle|$. We denote

$$\stackrel{\text{def}}{=} \overline{Span\{\mathcal{D}\}} \tag{7}$$

the closed linear span of the elements of $\mathcal{D}$. We denote $V^\perp$ the orthogonal complement of V in $\mathcal{H}$. We denote the orthogonal projection onto V and $V^\perp$ by P_V and $P_{V^\perp}$.

The sequences $(s_n)_{n \in \mathbb{N}}$, $(\gamma_n)_{n \in \mathbb{N}}$, $(R^n v)_{n \in \mathbb{N}}$ are always defined according to Table I. The coordinates $(\lambda_i)_{i \in I}$ are according to (5).

We also use the standard notations : $sgn(t) = 1$, if $t \geq 0$ and -1, if $t < 0$; $\#$ denotes the cardinal of a set; $\lfloor . \rfloor$ is the floor function.

E. Overview

In Section II, we define shrinkage, thresholding and gap functions. We also illustrate these definitions by several examples. In section III, we prove that as soon as θ is a shrinkage function: $(R^n v)_{n\in\mathbb{N}}$ converges and $\sum_{n\in\mathbb{N}} s_n\psi_{\gamma_n}$ exists. We also prove that $(s_n)_{n\in\mathbb{N}}$ is square summable. In Section IV, we prove that when θ is a thresholding function, $(s_n)_{n\in\mathbb{N}}$ is absolutely summable. This implies in particular that $(\lambda_i)_{i\in I}$ exists and is absolutely summable. In Section V, we prove that when θ is a gap function, the sequence $(s_n)_{n\in\mathbb{N}}$ is finite. Again, this implies that $(\lambda_i)_{i\in I}$ exists and is finite. Finally, in Section VI, we evaluate $\|\sum_{n\in\mathbb{N}} s_n\psi_{\gamma_n} - P_V v\|_{\mathcal{D}}$, when θ is a shrinkage function.

II. General shrinkage functions

A. Definitions

Definition 1: A function $\theta(\cdot) : \mathbb{R} \to \mathbb{R}$ is called a **shrinkage function** if and only if it satisfies:

1) $\theta(\cdot)$ is nondecreasing, i.e,

$$\forall t, t' \in \mathbb{R}, \quad t \leq t' \Longrightarrow \theta(t) \leq \theta(t');$$

2) $\theta(\cdot)$ shrinks the amplitude, i.e,

$$\forall t \in \mathbb{R}, \quad |\theta(t)| \leq |t|.$$

Notice that this implies

$$\theta(0) = 0,$$

and

$$\theta(-t) \leq 0 \leq \theta(t), \quad \forall t \geq 0. \tag{8}$$

Therefore, for any shrinkage function $\theta(\cdot)$ and any $t \in \mathbb{R}$, we know that:

$$\text{if } t \geq 0, \quad 0 \leq \theta(t) \leq t \quad \text{and } 0 \leq \theta(t)(t - \theta(t)),$$

$$\text{if } t \leq 0, \quad 0 \geq \theta(t) \geq t \quad \text{and } 0 \leq \theta(t)(t - \theta(t)).$$

As a conclusion,

$$\forall t \in \mathbb{R}, \quad \theta(t)(t - \theta(t)) \geq 0. \tag{9}$$

The inequality (8) also garantees that

$$\forall t \in \mathbb{R}, \quad |t|\ |\theta(t)| = t\theta(t). \tag{10}$$

Definition 2: Let $\theta(\cdot)$ be a shrinkage function, we call

- the **internal threshold:** $\tau^- \stackrel{\text{def}}{=} \inf_{t:\theta(t)\neq 0} |t|$
- the **external threshold:** $\tau^+ \stackrel{\text{def}}{=} \sup_{t:\theta(t)=0} |t|$.

Moreover, we say that $\theta(\cdot)$ is a **thresholding function** if and only if: $\tau^- > 0$, i.e.

$$\exists \tau > 0, \forall x \in \mathbb{R}, \qquad |x| \leq \tau \Rightarrow \theta(x) = 0. \tag{11}$$

If $\theta(\cdot)$ is a thresholding function, we trivially have

$$0 < \tau^- \leq \tau^+.$$

The internal and external thresholds are illustrated on Figure 1.

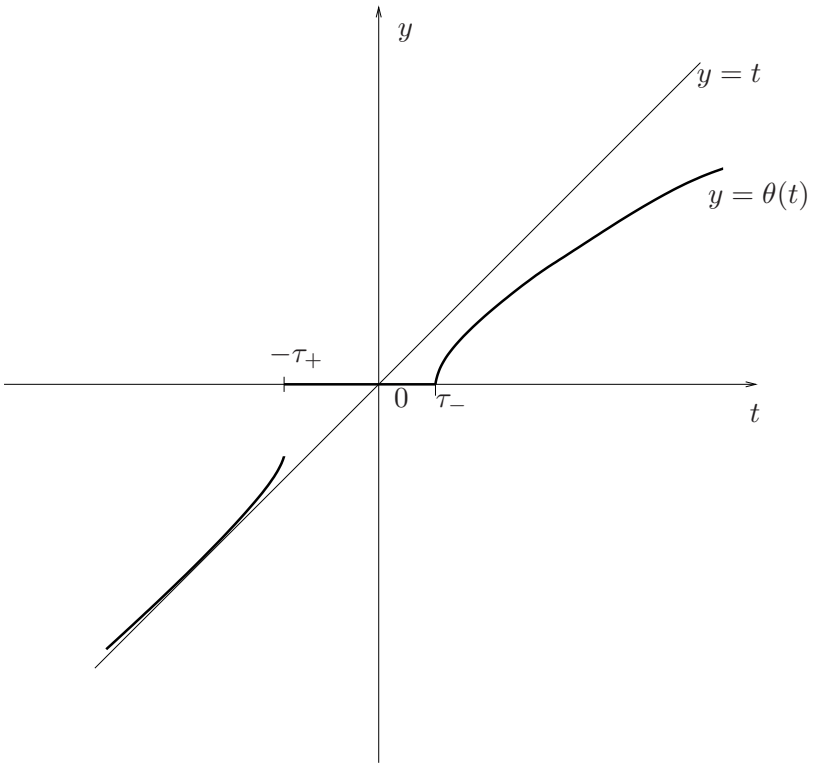

Fig. 1. Example of a thresholding function θ. It is non-gap. Its internal and external thresholds are not equal.

Since (9) holds for any shrinkage function, the following definition is valid.

Definition 3: The **gap** of a shrinkage function $\theta(\cdot)$ is defined by:

$$\text{gap}(\theta) \stackrel{\text{def}}{=} \inf_{t:\theta(t)\neq 0} \sqrt{\theta^2(t) + 2\theta(t)(t - \theta(t))}. \tag{12}$$

If $\text{gap}(\theta) > 0$, we call θ a gap shrinkage function and, if $\text{gap}(\theta) = 0$, the function is called a non-gap shrinkage function.

The following relation exists between the gap and the internal threshold of a shrinkage function. It proves in particular that any gap shrinkage function is a thresholding function.

Proposition 1: For any gap function $\theta(\cdot)$, we have

$$\mathrm{gap}(\theta) \leq \tau^-$$

where τ^- is the internal threshold of $\theta(\cdot)$.

Proof: The proof is given in Appendix.

B. Examples

Let us illustrate the above definitions through some examples.

1) For $\tau > 0$, the soft thresholding function $\rho_\tau(\cdot)$ defined by

$$\rho_\tau(t) = \mathrm{sgn}(t) \cdot \max(|t| - \tau, 0).$$

is a thresholding function and it is a non-gap shrinkage function, i.e., $\mathrm{gap}(\rho_\tau) = 0$.

2) For $\tau > 0$, the hard thresholding function defined by

$$h_\tau(t) = \begin{cases} t & \text{, if } |t| > \tau, \\ 0 & \text{, otherwise.} \end{cases}$$

is a thresholding function and it is a gap shrinkage function with gap τ.

3) The identity function defined as:

$$i(t) = t, \forall t \in \mathbb{R}, \tag{13}$$

is not a thresholding function and it is a non-gap shrinkage function.

4) For $\tau > 0$, the Non-Negative Garrote threshold function (see [21]) defined as:

$$\delta_\tau^G(t) = t \max\left(0, \left(1 - \frac{\tau^2}{t^2}\right)\right), \forall t \in \mathbb{R}, \tag{14}$$

is a thresholding function and it is non-gap.

5) For $0 < \tau_1 < \tau_2$, the firm shrinkage function (see [22]) defined as:

$$\delta_{\tau_1,\tau_2}(t) = \begin{cases} 0, & \text{if } |t| \leq \tau_1; \\ \mathrm{sgn}(t)\frac{\tau_2(|t|-\tau_1)}{\tau_2-\tau_1} & \text{if } \tau_1 < |t| < \tau_2; \\ t, & \text{if } |t| \geq \tau_2, \end{cases} \tag{15}$$

is a thresholding function and it is non-gap.

6) For $p \in \mathbb{N}$, $\tau > 0$, the generalized threshold function (see [23]) defined as:

$$\delta_\tau^p(t) = \begin{cases} t, & \text{if } |t| \leq \tau; \\ t - \frac{\tau^p}{t^{p-1}}(\mathrm{sgn}(t)^p), & \text{if } |t| > \tau, \end{cases} \tag{16}$$

a thresholding function and it is non-gap.

III. Convergence of the MP shrinkage for a shrinkage function

This section is devoted to prove that under mild condition, the MP shrinkage algorithm converges.

Proposition 2: Let $(\psi_i)_{i\in I}$ be a normed dictionary, $v \in \mathcal{H}$ and $\theta(\cdot)$ be a shrinkage function. For any $M > 0$ and any $v \in \mathcal{H}$, the quantities defined in Table I satisfy:

$$\|v\|^2 = \sum_{n=0}^{M-1} \left(s_n^2 + 2s_n(M_n - s_n)\right) + \|R^M v\|^2. \tag{17}$$

As a consequence, we have

$$\|v\|^2 \geq \sum_{n=0}^{M-1} s_n^2 + \|R^M v\|^2, \tag{18}$$

$$\sum_{n=0}^{+\infty} s_n^2 < +\infty, \tag{19}$$

$$\sum_{n=0}^{+\infty} |s_n|\ |M_n| < +\infty, \tag{20}$$

$$(\|R^n\|)_{n\in\mathbb{N}} \quad \text{is nonincreasing.} \tag{21}$$

Proof: We can deduce from

$$R^{n+1}v = R^n v - s_n \psi_{\gamma_n},$$

and $\langle \psi_{\gamma_n}, \psi_{\gamma_n} \rangle = 1$ that

$$\begin{aligned} \|R^{n+1}v\|^2 &= \|R^n v\|^2 - 2s_n\langle R^n v, \psi_{\gamma_n}\rangle + s_n^2 \\ &= \|R^n v\|^2 - 2s_n(M_n - s_n) - s_n^2. \end{aligned}$$

Summing these equalities for all $n = 0, \ldots, M-1$, we obtain after simplification

$$\|R^M v\|^2 = \|R^0 v\|^2 - \sum_{n=0}^{M-1} (s_n^2 + 2s_n(M_n - s_n)).$$

We then obtain (17) from $R^0 v = v$.

Using (9), we know that

$$s_n(M_n - s_n) = \theta(M_n)(M_n - \theta(M_n)) \geq 0.$$

Together with (17) this leads to (18).

Notice that this also provides (21). Moreover, (18) garantees that $(\sum_{n=0}^{M} s_n^2)_{M\in\mathbb{N}}$ is a bounded increasing sequence. It converges and (19) holds. We also have

$$\begin{aligned} 2\sum_{n=0}^{M-1} |s_n|\,|M_n| &= 2\sum_{n=0}^{M-1} s_n M_n \qquad \text{from (10)} \\ &= \|v\|^2 - \|R^M v\|^2 + \sum_{n=0}^{M-1} s_n^2 \qquad \text{from (17)} \\ &\leq \|v\|^2 + \sum_{n=0}^{+\infty} s_n^2. \end{aligned}$$

This ensures that (20) holds.

Now we can prove the convergence of the MP algorithm.

Theorem 1: Let $(\psi_i)_{i\in I}$ be a normed dictionary, $v \in \mathcal{H}$ and $\theta(\cdot)$ be a shrinkage function. The sequences defined in (4) satisfy:

$$(R^n v)_{n\in\mathbb{N}} \text{ converges.}$$

As a consequence,

$$\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n} \text{ exists.}$$

We denote the limit of $(R^n v)_{n\in\mathbb{N}}$ by $R^{+\infty}v$ and we trivially have

$$v = \sum_{n=0}^{+\infty} s_n \psi_{\gamma_n} + R^{+\infty} v.$$

Proof: The proof is based on Jones' proof for the convergence of projection pursuit regressions (see [24]) and the proof of Theorem 1 in [2].

First notice that the statement of the proposition is trivial for $v = 0$. We further assume that $v \neq 0$.

In order to prove the theorem, we prove that the sequence $(R^n v)_{n\in\mathbb{N}}$ is a Cauchy sequence. Before doing so, let us start with some preliminaries.

Notice first that for all $w_1, w_2 \in \mathcal{H}$, we have:

$$\begin{aligned} \|w_1 - w_2\|^2 &= \|w_1\|^2 - \|w_2\|^2 - 2\langle w_2, w_1 - w_2\rangle \\ &\leq \leq \|w_1\|^2 - \|w_2\|^2 + 2|\langle w_2, w_1 - w_2\rangle|. \end{aligned} \tag{22}$$

Moreover, for $N_2 > N_1 \geq 0$, from (6) we have

$$R^{N_1} v - R^{N_2} v = \sum_{n=N_1}^{N_2-1} s_n \psi_{\gamma_n}. \tag{23}$$

Finally, for any $n \geq 0$ and any $m \geq 0$,

$$\begin{aligned} |\langle R^m v, s_n \psi_{\gamma_n} \rangle| &= |s_n| \; |\langle \psi_{\gamma_n}, R^m v \rangle| \\ &\leq |s_n| \; \sup_{i \in I} |\langle \psi_i, R^m v \rangle| \\ &\leq \frac{1}{\alpha} |s_n| \; |M_m|. \end{aligned} \tag{24}$$

Let us now consider $N_2 > N_1 \geq 0$. Using (22), (23) and (24), we obtain

$$\begin{aligned} & \|R^{N_1} v - R^{N_2} v\|^2 \\ \leq \; & \|R^{N_1} v\|^2 - \|R^{N_2} v\|^2 + 2 |\langle R^{N_2} v, \sum_{n=N_1}^{N_2-1} s_n \psi_{\gamma_n} \rangle| \\ \leq \; & \|R^{N_1} v\|^2 - \|R^{N_2} v\|^2 + \frac{2}{\alpha} |M_{N_2}| \sum_{n=N_1}^{N_2-1} |s_n|. \end{aligned} \tag{25}$$

Using (21) of Proposition 2, we know that the sequence $(\|R^n v\|)_{n \in \mathbb{N}}$ is non-negative and non-increasing. Therefore, it converges to some value R_∞ and for any $\epsilon > 0$, there exits $K > 0$ such that for all $m > K$,

$$R_\infty^2 \leq \|R^m v\|^2 \leq R_\infty^2 + \epsilon^2.$$

As a consequence, for any $N_2 > N_1 \geq K$,

$$\|R^{N_1} v - R^{N_2} v\|^2 \leq \epsilon^2 + \frac{2}{\alpha} |M_{N_2}| \sum_{n=N_1}^{N_2} |s_n| \tag{26}$$

Using (20), we know that $\sum_{n=0}^{+\infty} |M_n||s_n| < +\infty$. Moreover, $0 \leq |s_n| \leq |M_n|$ for all $n \in \mathbb{N}$. So Lemma 2 (see Appendix) can be applied with $x_n \equiv |s_n|$ and $y_n \equiv |M_n|$. Two situations might occur :

- The first one is that: $\sum_{n=0}^{+\infty} |s_n| < +\infty$. In this case, we know that there is $K' > 0$ such that for any $N_2 > N_1 \geq K'$

  $$\sum_{n=N_1}^{N_2} |s_n| \leq \frac{\alpha}{2\|v\|} \epsilon^2.$$

  Moreover, from (17) we know that

  $$|M_{N_2}| = |\langle R^{N_2} v, \psi_{\gamma_{N_2}} \rangle| \leq \|R^{N_2} v\| \leq \|v\|.$$

  So (26) becomes : for any $\epsilon > 0$ there are K and $K' > 0$ such that for any $N_2 > N_1 \geq \max(K, K')$

  $$\|R^{N_1} v - R^{N_2} v\|^2 \leq \epsilon^2 + \epsilon^2.$$

  As a conclusion $(R^n v)_{n \in \mathbb{N}}$ is a Cauchy sequence.

- The second one is that: $\liminf_{q\to+\infty} |M_q| \sum_{n=0}^{q} |s_n| = 0$. In this case, let $\epsilon > 0$ and let $p > 0$ be an integer. We are going to estimate $\|R^m v - R^{m+p} v\|$, for $m > K$ (K is such that (26) holds). First, there is $q > m + p$ such that

$$|M_q| \sum_{n=0}^{q} |s_n| \leq \frac{\alpha}{2}\epsilon^2. \tag{27}$$

Moreover, we can decompose

$$\|R^m v - R^{m+p} v\| \leq \|R^m v - R^q v\| + \|R^{m+p} v - R^q v\|.$$

Applying (26) with $N_1 = m$ and $N_2 = q$ and using (27) we obtain

$$\|R^m v - R^q v\|^2 \leq \epsilon^2 + \epsilon^2.$$

Similarly, applying (26) for $N_1 = m + p$ and $N_2 = q$ and using (27) we obtain

$$\|R^{m+p} v - R^q v\|^2 \leq \epsilon^2 + \epsilon^2.$$

Hence, we finally obtain

$$\|R^m v - R^{m+p} v\| \leq 2\sqrt{2}\epsilon,$$

which proves that $(R^n v)_{n\in\mathbb{N}}$ is a Cauchy sequence.

As a conclusion, $(R^n v)_{n\in\mathbb{N}}$ converges. The second statement directly follows from (6).

Proposition 2 ensures that

$$\sum_{n=0}^{+\infty} |s_n|^2 \tag{28}$$

exists.

IV. l^1 NORM BOUNDS

In general, when $\mathcal{H}$ is an infinite dimensional space, we have no guarantee that

$$\sum_{n=0}^{+\infty} |s_n| \tag{29}$$

exists. A simple counter example consists in considering $(\psi_i)_{i\in I}$ a Riesz basis (for definition, see [25]) of $\mathcal{H}$, $v = \sum_{i\in I} s_i\psi_i \in \mathcal{H}$ such that $\sum_{i\in I} |s_i|$ diverges and $\theta(t) \equiv t$.

Below, we prove that (29) exists, whatever $v \in \mathcal{H}$ and whatever the dictionary, as soon as θ is a thresholding function.

Proposition 3: Let $(\psi_i)_{i\in I}$ be a normed dictionary, $v \in \mathcal{H}$ and $\theta(\cdot)$ be a thresholding function. The quantities defined in Table I satisfy:

$$\sum_{n=0}^{+\infty} |s_n| \leq \frac{\|v\|^2 - \|R^{+\infty}v\|^2}{\tau^-} \leq \frac{\|v\|^2}{\tau^-}, \tag{30}$$

where $\tau^- > 0$ denotes the internal threshold as defined in the Definition 2.

Proof: Let $M \in \mathbb{N}$ fixed. Using (18), we know that

$$\sum_{n=0}^{M-1} s_n^2 \leq \|v\|^2 - \|R^M v\|^2.$$

Together with (17), this leads to

$$\begin{aligned} \sum_{n=0}^{M-1} s_n M_n &= \frac{1}{2}\left(\|v\|^2 + \sum_{n=0}^{M-1} s_n^2 - \|R^M v\|^2\right) \\ &\leq \|v\|^2 - \|R^M v\|^2. \end{aligned}$$

Using (10) and the fact that $\theta(\cdot)$ is a thresholding function, for any $n \in \mathbb{N}$, we have:

$$s_n M_n = |s_n||M_n| \geq \tau^- |s_n|,$$

where the last inequality is obtained via the discussing on two cases: $s_n = 0$ or $s_n \neq 0$.

As a conclusion for all $M \in \mathbb{N}$ we have

$$\sum_{n=0}^{M-1} |s_n| \leq \frac{\|v\|^2 - \|R^M v\|^2}{\tau^-}. \tag{31}$$

Letting M go to infinity, we obtain (30).

Remark 1: The above upper bound does not depend on the dictionary $(\psi_i)_{i\in I}$. It holds for any $v \in \mathcal{H}$. We therefore do not expect this bound to be tight in any dedicated or applicative context.

Remark 2: As a side effect, the above proposition garantees that the coordinates λ_i exist for all $i \in I$ (see (5)). We even know that

$$\sum_{i\in I} |\lambda_i| < +\infty.$$

V. l^0 BOUNDS

If $\theta(\cdot)$ is a gap shrinkage function the MP shrinkage stops automatically after a finite number of iterations.

Proposition 4: Let $(\psi_i)_{i\in I}$ be a normed dictionary, $v \in \mathcal{H}$ and $\theta(\cdot)$ be a gap shrinkage function (i.e. $\mathrm{gap}(\theta) > 0$). The sequence $(s_n)_{n\in\mathbb{N}}$ defined in Table I satisfies:

$$\#\{n | s_n \neq 0\} \leq \lfloor \frac{\|v\|^2}{\mathrm{gap}(\theta)^2} \rfloor.$$

Proof: Suppose that the sequence $(s_n)_{n\in\mathbb{N}}$ contains M non-zero terms. Observing Definition 3, for each $s_n \neq 0$, we have:

$$s_n^2 + 2s_n(M_n - s_n) \geq \mathrm{gap}(\theta)^2,$$

where we recall that $M_n = \langle R^n v, \psi_{\gamma_n}\rangle$, $s_n = \theta(M_n)$.

From (17), we know that:

$$\|v\|^2 \geq \sum_{n\in\mathbb{N}:s_n\neq 0} \left(s_n^2 + 2s_n(M_n - s_n)\right) \geq M \cdot \mathrm{gap}(\theta)^2.$$

Noting that M is integer, we have:

$$M \leq \lfloor \frac{\|v\|^2}{\mathrm{gap}(\theta)^2} \rfloor.$$

Remark 3: An interesting consequence of the proposition is that

$$\#\{i \in I, \lambda_i \neq 0\} \leq \lfloor \frac{\|v\|^2}{\mathrm{gap}(\theta)^2} \rfloor.$$

In words, v is approximated with less than $\lfloor \frac{\|v\|^2}{\mathrm{gap}(\theta)^2} \rfloor$ non-zero coordinates.

VI. Bound on the residual error

In this section, we are interested in the residual error norm. The result concerns shrinkage functions. Before stating the result, let us give the following lemma:

Lemma 1: Let $(\psi_i)_{i\in I}$ be a normed dictionary, $v \in \mathcal{H}$ and $\theta(\cdot)$ be a shrinkage function. The sequence $(M_n)_{n\in\mathbb{N}}$ defined in Eq.(4) satisfies:

$$\limsup_{n\to+\infty} M_n \leq \sup_{t:\theta(t)=0} t, \tag{32}$$

and

$$\inf_{t:\theta(t)=0} t \leq \liminf_{n\to+\infty} M_n. \tag{33}$$

Proof: Let us prove the first statement. If $\sup_{t:\theta(t)=0} t = +\infty$ the statement is trivial. We therefore focus on the case $\sup_{t:\theta(t)=0} t < +\infty$. Let us assume that (32) does not hold. Then there exists $\epsilon > 0$ and an increasing sequence $(k_n)_{n\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$ such that

$$M_{k_n} \geq \sup_{t:\theta(t)=0} t + \epsilon, \quad \forall n \in \mathbb{N}.$$

So there exists an increasing sequence $(k_n)_{n\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$ such that

$$s_{k_n} = \theta(M_{k_n}) \geq \theta(\sup_{t:\theta(t)=0} t + \epsilon) > 0$$

This means that

$$\limsup_{n\to+\infty} s_n > 0.$$

The latter statement is impossible since, from (19), we know that $\lim_{n\to+\infty} s_n = 0$. This proves (32).

The proof of (33) is similar.

In particular, if the external threshold of $\theta(\cdot)$ is zero (i.e. $\tau^+ = 0$),

$$\lim_{n\to+\infty} M_n = 0,$$

since $\sup_{t:\theta(t)=0} t = \inf_{t:\theta(t)=0} t = 0$.

Recall that we have defined the semi-norm on $\mathcal{H}$ as

$$\|u\|_{\mathcal{D}} \stackrel{\text{def}}{=} \sup_{i\in I} |\langle u, \psi_i\rangle|, \quad \forall u \in \mathcal{H}.$$

Notice that $\|\cdot\|_{\mathcal{D}}$ is a norm as soon as $\mathcal{D}$ generates $\mathcal{H}$. Geometrically,

$$\{u \in \mathcal{H}, \|u\|_{\mathcal{D}} \leq \tau\}$$

is a polyhedron, for $\tau \geq 0$.

Recall that in (7) we denote $V \stackrel{\text{def}}{=} \overline{\text{Span}((\psi_i)_{i\in I})}$, the closure of vector space spanned by the dictionary $(\psi_i)_{i\in I}$, $V^\perp$ its orthogonal complement and we denote the orthogonal projection onto V and $V^\perp$ by P_V and $P_{V^\perp}$ respectively.

Proposition 5: Let $(\psi_i)_{i\in I}$ be a normed dictionary, $v \in \mathcal{H}$ and $\theta(\cdot)$ be a shrinkage function. The limits defined in Theorem 1 satisfy

$$\left\| \sum_{n=0}^{+\infty} s_n \psi_{\gamma_n} - P_V v \right\|_{\mathcal{D}} = \left\| R^{+\infty} v - P_{V^\perp} v \right\|_{\mathcal{D}} \leq \frac{\tau^+}{\alpha},$$

where τ^+ is the external threshold of $\theta(\cdot)$, as defined in Definition 2.

Proof: Let $\epsilon > 0$, from Lemma 1, we know that for any $k \geq 0$ there is $n_k \geq k$

$$\inf_{t:\theta(t)=0} t - \epsilon \leq M_{n_k} \leq \sup_{t:\theta(t)=0} t + \epsilon.$$

Given the definition of τ^+, we therefore know that

$$-\tau^+ - \epsilon \leq M_{n_k} \leq \tau^+ + \epsilon.$$

We rewrite

$$|M_{n_k}| \leq \tau^+ + \epsilon.$$

Moreover, since P_V is contractive and given the construction of M_{n_k}, we know that

$$|M_{n_k}| \geq \alpha \sup_{i\in I} |\langle R^{n_k} v, \psi_i\rangle| \geq \alpha \sup_{i\in I} |\langle P_V(R^{n_k} v), \psi_i\rangle|.$$

Therefore, for all $i \in I$,

$$|\langle P_V(R^{n_k} v), \psi_i\rangle| \leq \frac{\tau^+}{\alpha} + \frac{\epsilon}{\alpha}.$$

Since $(R^{n_k} v)_{k\in\mathbb{N}}$ converges to $R^{+\infty} v$ (see Theorem 1), we finally have

$$|\langle P_V(R^{+\infty} v), \psi_i\rangle| \leq \frac{\tau^+}{\alpha} + \frac{\epsilon}{\alpha},$$

for all $i \in I$. Since the above inequalities hold for any $\epsilon > 0$, we obtain

$$\left\|P_V(R^{+\infty} v)\right\|_{\mathcal{D}} \leq \frac{\tau^+}{\alpha}.$$

Moreover, using Theorem 1, we know that

$$P_{V^\perp}\left(R^{+\infty} v\right) \quad = \quad P_{V^\perp}(v) - P_{V^\perp}\left(\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n}\right) = P_{V^\perp}(v).$$

We therefore obtain

$$\left\|R^{+\infty} v - P_{V^\perp} v\right\|_{\mathcal{D}} = \left\|P_V(R^{+\infty} v)\right\|_{\mathcal{D}} \leq \frac{\tau^+}{\alpha}.$$

Using Theorem 1 (again), we also know that

$$\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n} \quad = \quad P_V\left(\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n}\right) = P_V(v) - P_V(R^{+\infty} v)$$

Therefore,

$$\left\|\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n} - P_V(v)\right\|_{\mathcal{D}} = \left\|P_V(R^{+\infty} v)\right\|_{\mathcal{D}} \leq \frac{\tau^+}{\alpha}.$$

This finishes the proof of the theorem.

Remark 4: A consequence of the above proposition is that when the MPS is used with a thresholding function, it provides a feasible point for the "Dantzig selector" (see [26]). The "Dantzig selector" consists in the optimization problem:

$$\min_{(\lambda_i)_{i\in I}} \sum |\lambda_i| \quad \text{subject to} \quad \|\sum_{i\in I} \lambda_i \psi_i - P_V v\|_{\mathcal{D}} \leq \frac{\tau^+}{\alpha}.$$

From Proposition 3, we know that the MPS provides a set of coordinates $(\lambda_i)_{i\in I}$ (see (5)) such that

$$\min_{(\lambda_i)_{i\in I}} \sum |\lambda_i|$$

is finite. Proposition 5 garantees that the constraint is satisfied.

ACKNOWLEDGEMENT

We would like to thank Prof. Alain Trouvé (ENS-Cachan), Prof. Michael Ng (HKBU) and Dr. Li Xiaolong (Peking Univ.) for the helpful discussions. We would also like to thank all the anonymous reviewers as their comments and suggestions contributed greatly to the quality of the paper.

APPENDIX

Proof of Proposition 1

Proof of $\mathrm{gap}(\theta) \leq \inf_{t:\theta(t)\neq 0} |t|$. Let $t_0 \in \mathbb{R}$ be such that $t_0 > \inf_{t:\theta(t)\neq 0} |t|$. We cannot simultaneously have $\theta(t_0) = 0$ and $\theta(-t_0) = 0$, since $\theta(\cdot)$ is nondecreasing. Let us denote

$$t = \begin{cases} t_0 & \text{, if } \theta(t_0) \neq 0 \\ -t_0 & \text{, if } \theta(t_0) = 0 \end{cases}$$

We have $\theta(t) \neq 0$ and given the definition of the gap, we know that

$$\begin{aligned} \mathrm{gap}(\theta)^2 &\leq \theta(t)^2 + 2\theta(t)(t - \theta(t)), \\ &= t^2 - (t - \theta(t))^2, \\ &\leq t^2 = t_0^2. \end{aligned}$$

As a conclusion, for any t_0 such that $t_0 > \inf_{t:\theta(t)\neq 0} |t|$, we have $\mathrm{gap}(\theta) \leq t_0$. So

$$\mathrm{gap}(\theta) \leq \inf_{t:\theta(t)\neq 0} |t|.$$

Lemma used in the proof of Theorem 1

This lemma is a variation on the Lemma used for the proof of Theorem 1 in [2].

Lemma 2: Let $(x_k)_{k\in\mathbb{N}}$ and $(y_k)_{k\in\mathbb{N}}$ be two sequences such that

$$\forall k \in \mathbb{N}, \quad 0 \leq x_k \leq y_k \tag{34}$$

and

$$\sum_{k=0}^{+\infty} x_k y_k < +\infty.$$

One of the following alternatives holds :

- either

$$\sum_{k=0}^{+\infty} x_k < +\infty$$

- or

$$\liminf_{j\to+\infty} y_j \sum_{k=0}^{j} x_k = 0.$$

Proof: First, since $(y_k)_{k\in\mathbb{N}}$ is a sequence of nonnegative real numbers, its inferior limit always exists. We

- either have $\liminf_{k\to+\infty} y_k > 0$,
- or $\liminf_{k\to+\infty} y_k = 0$

Let us first assume that

$$\liminf_{k\to+\infty} y_k > 0.$$

There exists $\epsilon > 0$ and $n > 0$ such that for any $k \geq n$, $y_k \geq \epsilon$. Therefore, we have

$$\epsilon \sum_{k=n}^{+\infty} x_k \leq \sum_{k=n}^{+\infty} x_k y_k < +\infty$$

and finally

$$\sum_{k=0}^{+\infty} x_k < +\infty.$$

The first alternative holds.

Let us from now on assume that

$$\liminf_{k\to+\infty} y_k = 0$$

and consider $\epsilon > 0$ and $m \geq 0$. Since $\sum_{k=0}^{+\infty} x_k y_k < +\infty$, there is $n \geq m$ such that

$$\sum_{k=n}^{+\infty} x_k y_k < \frac{\epsilon}{2}. \tag{35}$$

Since $\liminf_{k\to+\infty} y_k = 0$, there is $p \geq 0$ such that

$$y_{n+p} < \frac{1}{2\sum_{k=0}^{n-1} x_k}\epsilon. \tag{36}$$

Let $j \in \{n, \dots n+p\}$ be such that

$$y_j \leq y_k, \quad \forall k \in \{n, \dots n+p\}. \tag{37}$$

We have

$$\begin{aligned} y_j \sum_{k=0}^{j} x_k &= y_j \sum_{k=0}^{n-1} x_k + y_j \sum_{k=n}^{j} x_k \\ &\leq y_{n+p} \sum_{k=0}^{n-1} x_k + y_j \sum_{k=n}^{j} x_k \quad \text{from (37)} \\ &< \frac{\epsilon}{2} + \sum_{k=n}^{+\infty} x_k y_k \quad \text{from (36) and (34)} \\ &< \epsilon \quad \text{from (35).} \end{aligned}$$

As a conclusion, for any $\epsilon > 0$ and any $m \geq 0$, there is $j \geq m$ such that

$$y_j \sum_{k=0}^{j} x_k < \epsilon.$$

This means that the second alternative holds.

REFERENCES

[1] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constructive approximation*, vol. 13, no. 1, pp. 57–98, 1997.

[2] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[3] J. Friedman and W. Stuetzle, "Projection pursuit regression," *Journal of the American Statistical Association*, vol. 76, no. 376, pp. 817–823, Dec. 1981.

[4] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 255–260, Jan. 2006.

[5] R. DeVore and V. Temlyakov, "Some remarks on greedy algorithms," *Adv. Comput.Math.*, vol. 5, no. 1, pp. 173–187, 1996.

[6] V. Temlyakov, "Greedy algorithms and m-term approximation with regard to redundant dictionaries," *Journ. approx. Theory*, vol. 98, no. 1, pp. 117–145, 1999.

[7] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE, Trans. on Signal Processing*, vol. 51, no. 1, pp. 101–111, Jan. 2003.

[8] S. Krstulovic and R. Gribonval, "MPTK: Matching Pursuit made tractable," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'06)*, vol. 3, Toulouse, France, May 2006, pp. III–496 – III–499.

[9] D. Needell and J. Tropp, "Cosamp: Iterative recovery from incomplete and inaccurate samples," *Appl. Comp. Harmonic Anal.*, vol. 26, no. 3, pp. 301–321, 2009.

[10] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," 2008. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:0803.0811

[11] M. E. D. T. Blumensath, "Iterative hard thresholding for compressed sensing," *To appear in Applied and Computational Harmonic Analysis*, 2009.

[12] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit : Recursive function approximation with applications to wavelet decomposition," in *Proc. of 27th Asimolar Conf. on Signals, Systems and Computers*, Los Alamitos, 1993.

[13] S. Mallat, G. Davis, and Z. Zhang, "Adaptive time frequency decompositions," *SPIE, Journal of optical engineering*, vol. 33, no. 7, pp. 2183–2191, July 1994.

[14] S. Cotter, J. Adler, R. Rao, and K. Kreutz-Delgado, "Forward sequential algorithms for best basis selection," *IEE, proceedings in Vision, image and signal processing*, vol. 146, no. 5, pp. 235–244, Oct. 1999.

[15] M. E. D. T. Blumensath, "Gradient pursuits," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2370–2382, 2008.

[16] S. S. Chen, D. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.

[17] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math*, vol. 57, no. 11, pp. 1413–1457, 2004.

[18] P. L. Combettes and J.-C. Pesquet, "Proximal thresholding algorithm for minimization over orthonormal bases," *SIAM Journal on Optimization*, vol. 18, no. 4, pp. 1351–1376, October 2007.

[19] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *Annals of Appl. Stat.*, vol. 1, no. 2, pp. 302–332, 2007.

[20] T. Wu and K. Lange, "Coordinate descent algorithms for lasso penalized regression," *Annals of Appl. Stat.*, vol. 2, no. 1, pp. 224–244, 2008.

[21] H.-Y. Gao, "Wavelet shrinkage denoising using the non-negative garrote," *Journal of Computational and Graphical Statistics*, vol. 7, no. 4, pp. 469–488, 1998.

[22] H.-Y. Gao and A. G. Bruce, "Waveshrink with firm shrinkage," *Statistica Sinica*, vol. 7, pp. 855–874, 1997.

[23] Z. Zhao, "Wavelet shrinkage denoising by generalized threshold function," in *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, 18-21 August 2005.

[24] P. Huber, "Projection pursuit," *Ann. Statist.*, vol. 13, pp. 435–525, 1985.

[25] S. Mallat, *A Wavelet Tour of Signal Processing*. Boston: Academic Press, 1998.

[26] E. Candes and T. Tao, "The dantzig selector: Statistical estimation when p is much larger than n," *Annals of Stats.*, vol. 35, no. 6, pp. 2313–2351, 2007.